Towards General-Purpose Video Reconstruction through Synergy of Grid-Splicing Diffusion and Large Language Models

Jinliang Liu, Jianwei Zhang, Sen Yang, Jinxi Xiang, Xiyue Wang, Jieqiong Zhao, Zongxin Yang[†], Junhan Zhao[†]

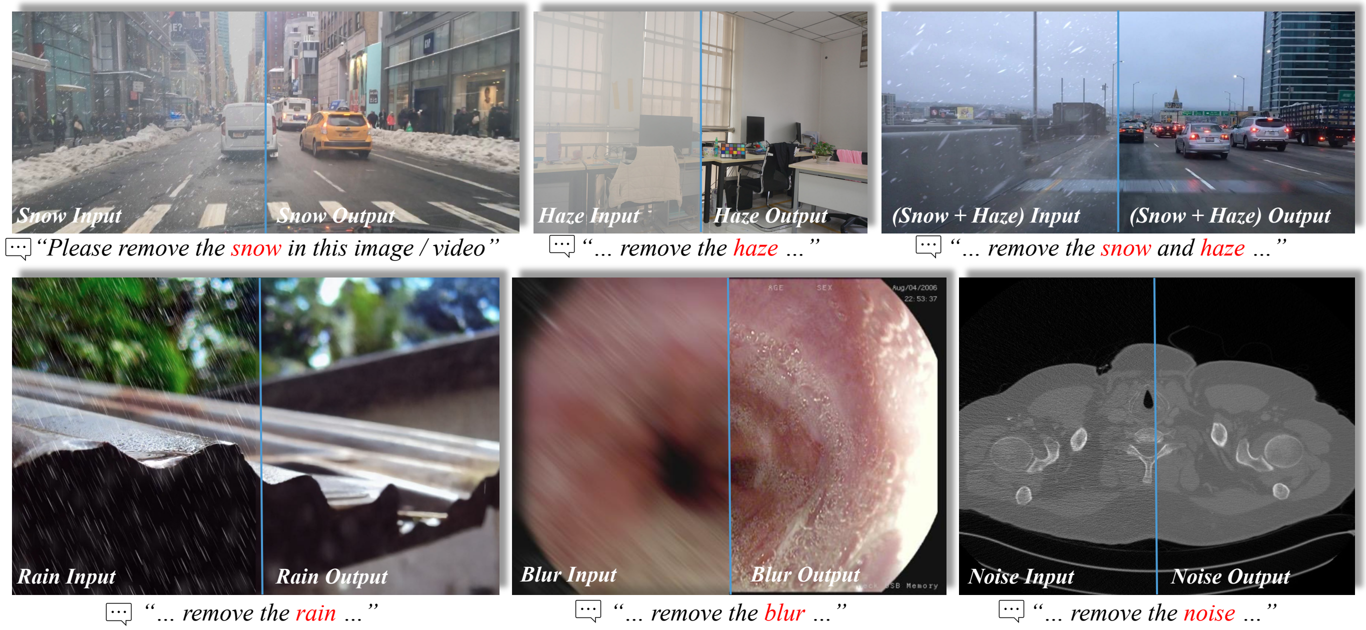

Snow Input
Snow Output
"Please remove the snow in this image / video"
Haze Input
Haze Output
"... remove the haze ..."
(Snow + Haze) Input
(Snow + Haze) Output
"... remove the snow and haze ..."
Rain Input
Rain Output
"... remove the rain ..."
Blur Input
Blur Output
AGE
SEX
Aug/04/2006
22:53:37
"... remove the blur ..."
Noise Input
Noise Output
"... remove the noise ..."

Fig. 1: Our GSDiff model demonstrates outstanding restoration results across tasks involving single degradation, real-world degradation, and mixed degradation. GSDiff effectively adapts to various degradation types through its innovative modules for temporal consistency and fine-grained detail preservation.

***Abstract*—Various forms of degradation, including noise, blur, and adverse weather conditions (e.g., rain, snow, and fog), significantly compromise video quality and system reliability across critical domains ranging from surveillance and medical imaging to entertainment. Previous research mainly focuses on network models tailored to specific degradation types, while recent unified frameworks and foundation models still face critical challenges in temporal consistency, automated degradation recognition, and detail preservation. Despite recent advances in foundation models, current approaches rely heavily on predefined degradation labels and remain focused on image-level operations, limiting their generalization to real-world scenarios and struggling with preserving fine-grained details. To address these challenges, we propose Grid Splicing Diffusion Model (GSDiff), a general framework for video reconstruction that leverages a novel grid splicing execution alongside instruction-tuned Large Language Model (LLM). GSDiff introduces three key innovative modules: (1) a LLM-driven degradation recognition module that enables automatic and fine-grained restoration guidance through zero-shot degradation analysis, (2) a Grid Splicing Module that organizes multiple frames into a unified grid structure to facilitate spatiotemporal feature processing, and (3) a Detail Preservation Module integrated with a Tail Refine Network to enhance fine-grained details during diffusion and post-processing. Extensive experiments demonstrate that GSDiff delivers state-of-the-art performance across a wide range of reconstruction tasks, including deraining, desnowing, denoising, and deblurring, propelling advancements in medical diagnostics and smart city applications.**

***Index Terms*—Unified Video Reconstruction, Large Language Models, Grid Splicing, Detail Preservation**

I. Introduction

In the digital era, video quality serves as a fundamental pillar of technological progress but remains susceptible to various degradations. From subtle sensor noise to the broader impact of adverse weather conditions, these impairments pose significant challenges across critical applications, including medical diagnostics, surveillance systems, and entertainment platforms, where precision and reliability are crucial. Video reconstruction algorithms play a vital role in restoring de-

Jinliang Liu is with Centre for Artificial Intelligence, University of Technology Sydney, 15 Broadway, Ultimo NSW 2007, Australia. E-mail: jinliang.liu@student.uts.edu.au. Jianwei Zhang is with Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA, 90089, USA. E-mail:jzhang17@usc.edu. Sen Yang, Jinxi Xiang and Xiyue Wang are with the Department of Radiation Oncology, Stanford Medicine, CA, USA. Email: (seny, xiangjx, xiyue)@stanford.edu. Jieqiong Zhao is with the School of Computer and Cyber Sciences at Augusta University. Email:jiezhao@augusta.edu. Zongxin Yang and Junhan Zhao are with Harvard University, 25 Shattuck Street, Boston, MA 02115, USA. Email: (zongxin_yang, junhan_zhao)@hms.harvard.edu.

[†] Correspondences to Z.Yang and J.Zhao.

Manuscript received April 19, 2021; revised August 16, 2021.

graded footage, enhancing both human visual perception and the performance of downstream deep learning models in video analysis tasks [1]–[6]. Traditional approaches address a single type of degradation through task-specific neural architectures. Single-task models achieve strong performance on individual tasks such as deraining [7], [8], desnowing [9], [10], denoising [11], [12] and deblurring [13], [14] under controlled settings. However, such specialized architectures exhibit limited generalization capabilities to real-world scenarios where multiple types of degradation appear simultaneously.

To address these generalization challenges, unified frameworks emerge as a solution to handle multiple types of degradation simultaneously [15]–[22]. Such multitask architectures demonstrate enhanced generalization capabilities across different degradation patterns, eliminating the need for training separate models for individual degradation types. However, unified approaches heavily rely on paired synthetic training data, which leads to significant domain gaps between synthetic and real-world scenarios. This **synthetic-to-real gap** substantially limits model performance when encountered unseen degradation patterns in real-world applications. Recent advances in foundation models introduce a promising direction for image reconstruction [23]. By leveraging large-scale pre-trained models, current approaches demonstrate enhanced generalization across diverse degradation patterns. Integration of language models such as CLIP [24] enables text-guided reconstruction, where natural language instructions direct the restoration process [25]. This approach not only improves model generalization but also provides flexible control through linguistic supervision, allowing human-in-the-loop specification of reconstruction objectives.

Despite progress in text-guided reconstruction methods, existing video processing techniques still face several challenges. ***1) Limited flexibility and accuracy in automated degradation recognition.*** The direct application of visual-language features lacks the capability to automatically recognize and adapt to diverse degradation scenarios. Moreover, without sophisticated language understanding capabilities, this approach struggles to establish precise mappings between visual degradation patterns and their semantic descriptions. ***2) Existing methods ignore temporal dependencies across video frames.*** Such frame-independent reconstruction paradigm inherently suffers from temporal inconsistency and flickering artifacts that significantly degrade the visual quality of long sequences. ***3) The variational autoencoder (VAE) architecture compromises fine details and textures.*** Although the VAE structure used in models such as Stable Diffusion effectively removes degradation, it often sacrifices high-frequency details and features, reducing sharpness and overall fidelity in reconstructed frames.

To address these challenges, we propose a novel video reconstruction framework, GSDiff (Grid Splicing Diffusion), which incorporates three key innovative modules. To tackle challenge ***1)***, we propose a LLM-driven zero-shot degradation recognition system that enables automatic and fine-grained restoration guidance without requiring explicit degradation annotations. We employ InstructBLIP to perform single-frame degradation analysis, leveraging its instruction-tuning capabilities to generate comprehensive and task-relevant degradation descriptions. We design a Degradation Extractor module that effectively integrates these semantic descriptions with CLIP visual features, establishing precise mappings between visual degradation patterns and their semantic characterizations. To tackle challenge ***2)***, we introduce the Grid Splicing module that arranges multiple frames into a unified grid structure. Through the self-attention mechanism in U-Net [26], this design enables simultaneous processing of spatiotemporal features, ensuring temporal consistency while enhancing reconstruction details through cross-frame information sharing [27], and significantly reducing computational overhead. To resolve challenge ***3)***, we introduce a Detail Preservation Module (DPM) coupled with a Tail Refine Network for high-fidelity detail reconstruction. The proposed dual-module design synergistically combines multi-stage attention mechanisms with hierarchical feature processing. DPM leverages temporal-guided attention and degradation-aware modulation to preserve fine-grained details, while the Tail Refine Network employs a symmetric encoder-decoder architecture with window-based attention mechanisms to enhance local-global feature representations. This architecture effectively mitigates the detail loss inherent in VAE-based frameworks [28] while maintaining temporal consistency. Extensive experiments demonstrate that GSDiff achieves state-of-the-art performance across four representative video reconstruction tasks (deraining, desnowing, denoising, and deblurring). Notably, our method overperforms SOTA models on both synthetic benchmark datasets and real-world degraded videos, demonstrating remarkable generalization capabilities and producing visually compelling reconstruction results (Shown in Fig. 1). Our contributions include:

- We design a zero-shot degradation recognition system that combines LLM-driven analysis with CLIP visual features. This system leverages InstructBLIP for automatic degradation analysis and introduces a Degradation Extractor module for precise visual-semantic mapping, eliminating the need for explicit degradation annotations.
- We propose GSDiff, integrating a novel Grid Splicing module with diffusion models for video reconstruction. The Grid Splicing technique enables efficient spatiotemporal feature processing through a unified grid structure, significantly reducing computational overhead while maintaining temporal consistency.
- We introduce a dual-stage detail enhancement framework that combines Detail Preservation Module (DPM) with Tail Refine Network (TRN), achieving state-of-the-art performance in preserving fine-grained details across diverse restoration tasks, particularly in medical imaging applications.

II. Related Work

A. Unified Low-level Restoration

Traditional image and video restoration methodologies concentrate on addressing singular degradation types, including denoising, deblurring, deraining, or desnowing in isolation [7]–[14]. Despite their effectiveness for specific tasks, these approaches require maintaining separate models, increasing computational overhead and limiting applicability.

Motivated by tackling complex, real-world degradations and the insight that diverse degradation types exhibit shared underlying patterns, the research community demonstrates a paradigm shift towards unified restoration frameworks. These frameworks, which leverage joint feature learning across multiple tasks, continue to yield substantial advancements. Li *et al.* [29] introduce AirNet, which integrates a Contrastive-Based Degraded Encoder with a Degradation-Guided Restoration Network, establishing fundamental principles for unified restoration of multiple unknown degradation types. Extending this foundation, Kulkarni *et al.* [22] investigate lightweight architectural designs, presenting UVRNet with a dual-stream architecture that demonstrates competitive performance in rain and snow fog removal while utilizing minimal parameters. Concurrent with the emergence of Transformers in computer vision, TransWeather [18] adopts this architectural paradigm for weather degradation restoration. It effectively removes multiple types of weather degradations using sophisticated intra-patch attention mechanisms and learnable weather type embeddings. To enhance the generalization capabilities of models, Chen *et al.* [21] propose a novel knowledge-learning paradigm that incorporates two-stage knowledge learning with multi-contrastive regularization, enabling unified pre-trained weights to address multiple weather degradations simultaneously. PromptIR [30] advances the field by incorporating prompt learning methodologies, encoding degradation-specific information as prompts to systematically guide the restoration network, representing a substantial advancement towards more adaptable and universal restoration approaches. ViWS-Net by Yang *et al.* [31] demonstrates multi-weather removal capabilities in video restoration through weather-invariant video transformer encoders, temporally-aware weather messenger mechanisms, and adversarial backpropagation-based weather feature suppression. Nevertheless, compared to image-based methodologies, video restoration architectures remain relatively understudied. Motivated by this research gap, we concentrate our investigation on video-level reconstruction tasks, aiming to advance the development of robust and efficient unified restoration frameworks in the video domain.

B. Large Language Models

Large Language Models (LLMs) demonstrate revolutionary advances in vision-language multimodal understanding and generation tasks [32]–[38]. These models excel in not only traditional tasks such as text generation and reasoning,, but also cross-modal scenarios including visual question answering and scene understanding. The emergence of GPT-3 [32] demonstrates strong zero-shot and few-shot transfer capabilities through pre-training with 175 billion parameters, establishing a foundation for subsequent multimodal tasks. Building upon this success, PaLM [34] extends the model scale to 540 billion parameters and achieves breakthroughs in complex reasoning through the Pathways system. BLIP [36] significantly improves vision-language alignment through innovative image-text pair generation strategies and cross-modal filtering mechanisms, establishing new benchmarks in image captioning and visual question answering. LLaMA [35] introduces compute-efficient open-source models that achieve comparable performance to proprietary systems. LM4LV [39] pioneers the application of frozen LLMs in low-level vision tasks, demonstrating that LLMs can effectively handle fundamental computer vision problems without requiring training using multi-modal data or prior knowledge. X-Former [40] integrates contrastive learning with masked image modeling to enhance multimodal large language models (MLLMs) for visual understanding [41], achieving state-of-the-art performance in tasks that demand detailed visual comprehension. GPT-4 [33] demonstrates unprecedented multimodal understanding capabilities in commercial applications. Notably, InstructBLIP [37] incorporates instruction learning paradigms into vision-language models, enabling more precise visual understanding and description capabilities. This instruction-tuning capability allows accurate identification and description of various degradation patterns, providing critical support for our zero-shot degradation recognition system. Specifically, we leverage InstructBLIP to generate detailed degradation descriptions and effectively fuse this semantic information with CLIP visual features for precise degradation pattern modeling. Despite recent advances, leveraging LLMs for video reconstruction presents both exciting opportunities and significant challenges. By integrating the visual understanding and semantic expression capabilities of LLMs, video reconstruction tasks could see improvements in degradation recognition, temporal modeling, and high-fidelity restoration.

III. Method

A. Our approach

*1) **LLM-driven Degradation Guidance**:* Traditional video restoration methods rely on explicit degradation annotations, limiting their real-world generalization. The proposed zero-shot framework leverages Large Language Models and CLIP representations for automatic degradation recognition and restoration guidance. To maintain the robust pre-trained knowledge and ensure efficient training, both the LLM, CLIP and U-net from stable diffusion [26] are kept frozen during training, with only the Degradation Extractor being optimized.

Automatic Degradation Identify via LLM: In the automatic degradation analysis module, we introduce InstructBLIP [37], a Large Language Model (LLM) with powerful visual understanding capabilities, as the degradation condition recognizer. Compared to traditional degradation analysis methods, the LLM provides more fine-grained and semantically rich degradation feature descriptions through accurate interpretation of complex visual degradation phenomena. Due to the temporal consistency of degradation information in video sequences V_d, we only feed the first frame F_1 with a pre-designed prompt `"What kinds of degradation types are in this image?"` into the model, ensuring analysis accuracy while significantly improving computational efficiency. The fully automated analysis process generates detailed degradation type descriptions (*e.g.*, `"heavy rain with motion blur"`, `"dense snow with haze"`), which demonstrate InstructBLIP's precise recognition of complex weather and motion degradations. These descriptions later are fed into the CLIP's text encoder for subsequent processing guidance.

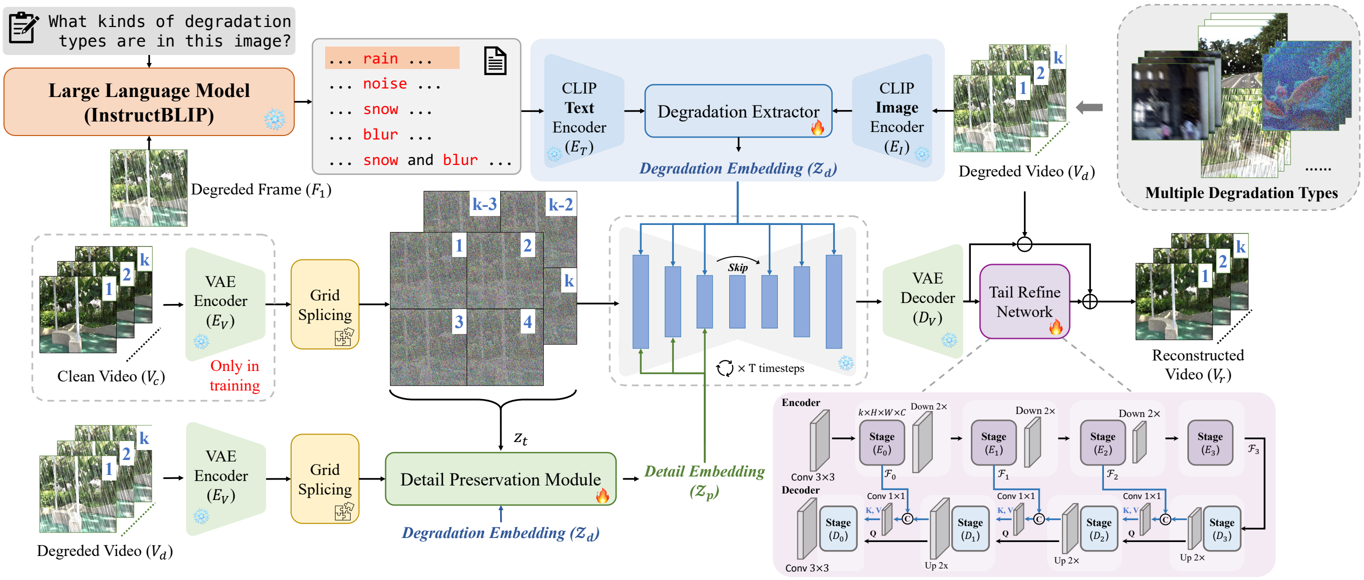


Fig. 2: The overall architecture of GSDiff for video reconstruction. Our framework consists of three key components: (1) a LLM-driven degradation recognition system using InstructBLIP and Degradation Extractor for automatic degradation analysis, (2) a Grid Splicing module that arranges multiple frames into a unified grid structure for spatiotemporal feature processing, and (3) a Detail Preservation Module coupled with Tail Refine Network for high-fidelity reconstruction. The degraded video (V_d) is used for both training and inference, while the clean video (V_c) is only required during training.

CLIP-based Dual Branch Guidance: We design the Dual Branch Guidance (DBG) of pre-trained CLIP model to extract complementary representations from both visual and textual domains. The image encoder E_I processes the k-frame degraded video sequence V_d and aggregates features to obtain a global representation in the image domain:

$$\mathcal{F}_I = \sum_{t=1}^{k} \frac{E_I(V_d^t)}{k}, \quad V_d^t \in \mathbb{R}^{H \times W \times 3}, \tag{1}$$

where $\mathcal{F}_I \in \mathbb{R}^d$ denotes the aggregated visual features with dimension d. For text feature extraction, we design a structured prompt template "`Please remove the` [*degradation*] `from this image or video`", where the degradation placeholder is filled with specific types identified by LLM (*e.g.*, "`heavy rain with motion blur`", "`dense snow with haze`"). The text encoder E_T maps the processed prompt into feature representation $\mathcal{F}_T \in \mathbb{R}^d$ with matching dimension for feature fusion.

Degradation Extractor: The Degradation Extractor (DE) module fuses visual features $\mathcal{F}_I$ and textual features $\mathcal{F}_T$ through feature concatenation, which enables the module to capture both visual degradation patterns from video frames and semantic-level degradation descriptions from LLM. The degradation embedding is generated through:

$$\mathcal{Z}_d = \mathrm{MLP}([\mathcal{F}_I; \mathcal{F}_T]) \tag{2}$$

where the degradation embedding $\mathcal{Z}_d$ is injected into multiple scales of U-Net to provide fine-grained degradation guidance for video restoration.

*2) **Grid Splicing**:* The Grid Splicing method originates from the exploration of the need for temporal consistency in video reconstruction tasks. By arranging video frames in a grid structure, the model can effectively capture and preserve the temporal dependencies between frames. In traditional frame-by-frame video reconstruction methods, the lack of temporal coordination often leads to visual artifacts (such as flickering between frames and inconsistent details), which affects the overall quality of reconstruction. To address this problem, we propose GSDiff, which uses a Grid Splicing structure to enhance temporal consistency, making the reconstructed video frames visually smoother and more high-fidelity. Taking a heuristic from daily experience, imagine a comic strip where multiple panels display sequential moments in time in one sheet. Instead of processing each "panel" (video frame) in isolation, our Grid Splicing approach arranges these sequential frames in a structured layout, allowing the model to recognize temporal relationships at a glance. This enhances visual consistency while reducing computational complexity.

Given an input video sequence $V = \{f_t\}_{t=1}^{k}$ containing k frames, each frame $f_t \in \mathbb{R}^{H \times W \times 3}$ represents an RGB image. We arrange these frames into a grid $\mathcal{G}_{n,m}$ with n rows and m columns:

$$\mathcal{G}_{n,m} : \mathbb{R}^{k \times H \times W \times 3} \rightarrow \mathbb{R}^{(H \cdot n) \times (W \cdot m) \times 3}, \tag{3}$$

where $k = n \times m$ represents the total number of frames, each grid cell $g_{i,j} \in \mathbb{R}^{H \times W \times 3}$ corresponds to a frame f_t, satisfying:

$$t = i{\cdot}m{+}j{+}1, \quad i \in \{0, 1, \ldots, n{-}1\}, \quad j \in \{0, 1, \ldots, m{-}1\}. \tag{4}$$

This grid representation not only reduces the memory overhead of a single processing but also allows the model to exploit the spatial arrangement in the grid to extract spatiotemporal features between frames. When implementing frame reorganization, the conversion from video to grid is defined as follows:

$$V \rightarrow \mathcal{G}_{n,m} = \{G_l | G_l = f_l\}_{l=1}^{k}. \tag{5}$$

Through GSDiff, the video frame is first encoded into a low-dimensional feature representation through the VAE encoder, and then input into the U-Net of the diffusion model according to the grid arrangement. In the self-attention mechanism of U-Net, the model can not only process the spatial information of each frame but also capture the temporal information between frames through the grid structure. In this way, the model can identify and eliminate degradation phenomena (such as rain, snow, fog, etc.) in the video during the reconstruction process, making the reconstructed video frames more coherent.

*3) **Detail Preservation Module**:* The Detail Preservation Module (DPM) preserves fine textures often lost during video restoration. It ensures the restoration precision via analyzing both the degraded content and the specific type of degradation. To achieve efficient detail reconstruction, DPM integrates three key information sources: degraded video content features $f_d = \mathrm{GS}(\mathrm{E}_V(V_d)) \in \mathbb{R}^{C\times H\times W}$, diffusion timestep features z_t, and degradation embedding $\mathcal{Z}_d$. Specifically, degraded video features carry structural information, timestep features provide denoising process guidance, and degradation embedding ensures restoration optimization for specific degradation types. To effectively process multi-scale detail information, DPM employs a four-stage encoder-decoder architecture. For feature fusion, we design a two-stage attention mechanism:

$$F_{\text{temp}} = \mathrm{Attn}(Q = z_t, K = f_d, V = f_d), \tag{6}$$

$$F_i = \mathrm{CrossAttn}(Q = F_{\text{temp}}, K = \mathcal{Z}_d, V = \mathcal{Z}_d). \tag{7}$$

The first-stage attention leverages timestep information z_t to guide degraded feature enhancement, while the second stage incorporates degradation type information for adaptive detail reconstruction. The generated multi-scale features $\mathcal{Z}_p = \{F_i\}_{i=0}^{3}$ are injected as residual information into corresponding diffusion U-Net layers, enabling progressive detail enhancement. Parameters are updated through zero-initialized 1×1 convolutions, maintaining pretrained features while achieving task-specific precise modulation. This design significantly improves detail reconstruction quality through multi-source information fusion and multi-scale residual injection, without compromising the original network performance.

*4) **Tail Refine Network**:* The Tail Refine Network (TRN) employs a symmetric encoder-decoder architecture with four stages E_i and D_i ($i \in \{0, 1, 2, 3\}$) to process n consecutive frames. Starting with a 3×3 convolution to extract C-channel features ($n \times H \times W \times C$), each stage integrates Swin Transformer-based attention inspired by [42]. In the encoder part, each stage (except E_3) progressively downsamples features through Down 2x operations, generating multi-scale representations $\mathcal{F}_i$ that connect to corresponding decoder stages through skip connections. The encoder stages form a hierarchical feature extraction pipeline, capturing both local and global features at different spatial resolutions.

During decoding, each stage D_i (except D_3) incorporates multiple TRN Swin Transformer Decoder Blocks, as shown in Fig. 3. For each block, features are processed through a Conv 3×3 layer and upsampled from the previous stage through Up 2× operations as queries **Q**, while encoder features $\mathcal{F}_i$ are mapped to key-value pairs (**K**, **V**) through 1×1 convolution.

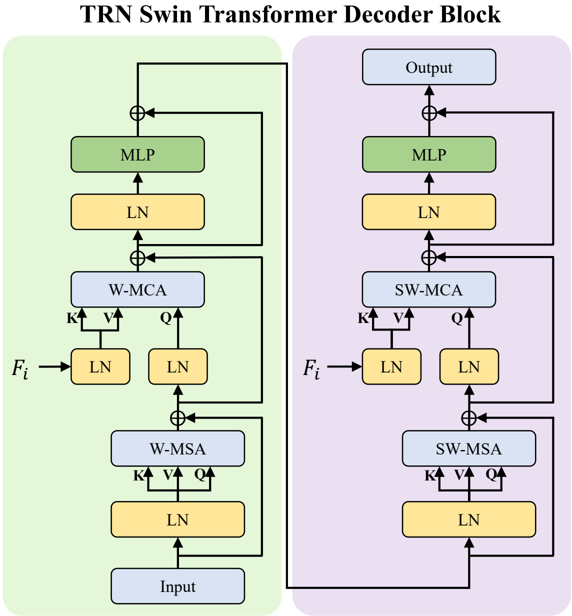


Fig. 3: Architecture of the Tail Refine Network (TRN)'s Swin Transformer Decoder Block. This block features dual parallel branches, each with MLP layers and a Layer Normalization (LN), followed by Window Multi-head Self-Attention (W-MSA) and Shifted Window Multi-head Self-Attention (SW-MSA) modules for effective feature extraction.

The decoder block consists of a dual-branch architecture with alternating window attention mechanisms - W-MSA/W-MCA in one branch and SW-MSA/SW-MCA in the other. Each attention module is followed by Layer Normalization (LN) and Multi-Layer Perceptron (MLP) with residual connections (⊕). By stacking multiple blocks in each decoder stage, our model progressively refines features via hierarchical attention, effectively capturing local details and global context.

The final output is obtained by: $V_r = (D_V(V_d) - V_d) + T(V_d)$, where $D_V(V_d)$ and $T(V_d)$ denote VAE Decoder and Tail Refine Network outputs respectively. The subtraction operation helps remove the degraded components from VAE Decoder output before combining with the refined details from Tail Refine Network. This design enables the network to focus on detail enhancement while maintaining structural consistency through the complementary combination of VAE reconstruction and refined features.

*5) **Loss Function**:* We design three complementary loss terms to optimize our framework. First, we introduce a conditional diffusion loss that incorporates degradation information into the denoising process:

$$\mathcal{L}_{diff} = \mathbb{E}_{G_c,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(G_c^t, t, \mathcal{Z}_d)\|_2^2\right], \tag{8}$$

where G_c^t denotes the clean grid structure with added noise at timestep t, and $\mathcal{Z}_d \in \mathbb{R}^d$ represents the degradation embedding extracted by our Dual-branch Degradation Guidance module. To ensure high-fidelity reconstruction, we employ an L1 reconstruction loss:

$$\mathcal{L}_{rec} = \mathbb{E}_{V_c,V_r}\left[\|V_c - V_r\|_1\right] \tag{9}$$

TABLE I: Quantitative comparison with state-of-the-art methods on four restoration tasks. The proposed GSDiff consistently outperforms existing approaches across all tasks, achieving superior performance in terms of both PSNR (dB) and SSIM metrics.

Method	Type	Source	Restoration Tasks									
			Derain		Desnow		Denoise		Deblur		Average	
			PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
All-in-One [43]	Image	CVPR'20	24.33	0.8217	29.13	0.8901	24.39	0.8077	23.22	0.7937	25.27	0.8283
AirNet [29]	Image	CVPR'22	26.61	0.8949	31.15	0.9410	28.09	0.8233	25.76	0.8120	27.90	0.8678
TransWeather [18]	Image	CVPR'22	25.22	0.9118	30.17	0.9325	28.87	0.8313	24.35	0.7852	27.15	0.8652
Restormer [44]	Image	CVPR'22	26.79	0.9241	32.30	0.9531	30.03	0.8323	25.12	0.7805	28.56	0.8725
PromptIR [21]	Image	NIPS'23	26.85	0.9250	32.47	0.9568	31.01	0.8335	25.42	0.7833	28.94	0.8747
WeatherDiffusion [45]	Image	TPAMI'23	25.36	0.9123	33.10	0.9442	30.40	0.9113	25.12	0.8193	28.50	0.8968
WGWS-Net [46]	Image	CVPR'23	27.64	0.9210	32.89	0.9313	31.58	0.9328	26.31	0.9065	29.61	0.9229
ViWS-Net [31]	Video	ICCV'23	28.52	0.9332	34.23	0.9441	32.94	0.9516	28.17	0.9194	30.97	0.9371
GSDiff (ours)	Video	— —	**30.15**	**0.9491**	**35.22**	**0.9518**	**33.71**	**0.9722**	**30.24**	**0.9207**	**32.33**	**0.9485**

where V_c and V_r represent the clean and reconstructed sequence. Additionally, we introduce a contrastive loss to enhance degradation-specific feature learning:

$$\mathcal{L}_{con} = \mathbb{E}_{V_{d1},V_{d2}} \left[\|Z_{d1} - Z_{d2}\|_2^2 - \|Z_{d1} - Z_{d1}^{+}\|_2^2 \right], \quad (10)$$

where Z_{d1} and Z_{d2} are embeddings of different degradation types, and Z_{d1}^{+} represents the embedding from a temporally adjacent frame. The total loss combines these terms with balanced weights:

$$\mathcal{L}_{total} = \lambda_{diff}\mathcal{L}_{diff} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{con}\mathcal{L}_{con}, \quad (11)$$

where we set $\lambda_{diff} = 1.0$, $\lambda_{rec} = 0.2$, and $\lambda_{con} = 0.01$ in our implementation.

IV. Experiments Settings

Datasets: This study utilizes four representative datasets for video and image processing tasks. RainMotion [47], synthesized from the NTURain framework, incorporates five large-scale rain streak masks with natural motion trajectories, effectively simulating real-world rainfall scenarios. The KITTI-snow dataset [48], designed for outdoor video desnowing, comprises 50 video sequences with diverse snowflake properties and Gaussian blur, enhancing the task complexity. For denoising, the DAVIS dataset [49] follows FastDVDnet's preparation paradigm [50], where clean patches are randomly sampled and noisy counterparts are generated using Additive White Gaussian Noise (AWGN) with standard deviations of 30 and 50. The GoPro dataset [51], targeting image deblurring, contains 3,214 high-resolution (1,280×720) frames split into 2,103 training and 1,111 testing samples, featuring precise correspondence between real-world blurred scenes and their ground truth captured via high-speed cameras. These datasets provide rigorous evaluation benchmarks for deraining, desnowing, denoising, and deblurring tasks.

Evaluation Metrics: For quantitative evaluation, we adopt two standard metrics: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). PSNR measures pixel-level reconstruction accuracy through mean squared error in logarithmic scale, while SSIM evaluates structural similarity by considering luminance, contrast, and structural information.

A. Quantitative comparison

To comprehensively evaluate the effectiveness of our proposed method, we conduct extensive comparisons against several state-of-the-art approaches. Specifically, we select the classical all-weather image restoration methods including All-in-One [43], AirNet [29] and TransWeather [18] as our baselines. Furthermore, we compare with recent prominent restoration frameworks that demonstrate superior performance on continuous degradation tasks, namely Restormer [44] and PromptIR [21]. To ensure a thorough comparison with current advances, we also include latest diffusion-based approaches, WeatherDiffusion [45] and WGWS-Net [46], which have shown remarkable capacity in weather removal tasks. Additionally, we include ViWS-Net [31], a recent video-based weather removal method, in our analysis.

The quantitative results in Tab. I highlight the superiority of GSDiff across all restoration tasks, with consistent improvements over state-of-the-art methods. For deraining, GSDiff achieves 30.15 dB in PSNR and 0.9491 in SSIM, outperforming ViWS-Net by 1.63 dB and 0.0159, respectively. In the desnowing task, GSDiff reaches 35.22 dB PSNR and 0.9518 SSIM, surpassing ViWS-Net by 0.99 dB and 0.0077. Similarly, in denoising and deblurring, GSDiff attains 33.71 dB and 30.24 dB in PSNR, with SSIM scores of 0.9722 and 0.9207, consistently leading over other methods. On average, GSDiff achieves 32.33 dB PSNR and 0.9485 SSIM, significantly exceeding the baseline ViWS-Net (30.97 dB / 0.9371). By leveraging temporal information, our method outperforms image-based approaches like WGWS-Net and PromptIR, delivering more robust and superior results across diverse weather degradation scenarios. In the comparison results, the top-performing method is highlighted in " pink ", and the second-best method is highlighted in " blue ".

B. Qualitative Comparison

We compare our proposed method GSDiff with the state-of-the-art video restoration method ViWS-Net [31], the classic image restoration method Restormer [44], and the diffusion-based method WeatherDiffusion [45]. As shown in Fig. 4, our method demonstrates superior restoration performance across

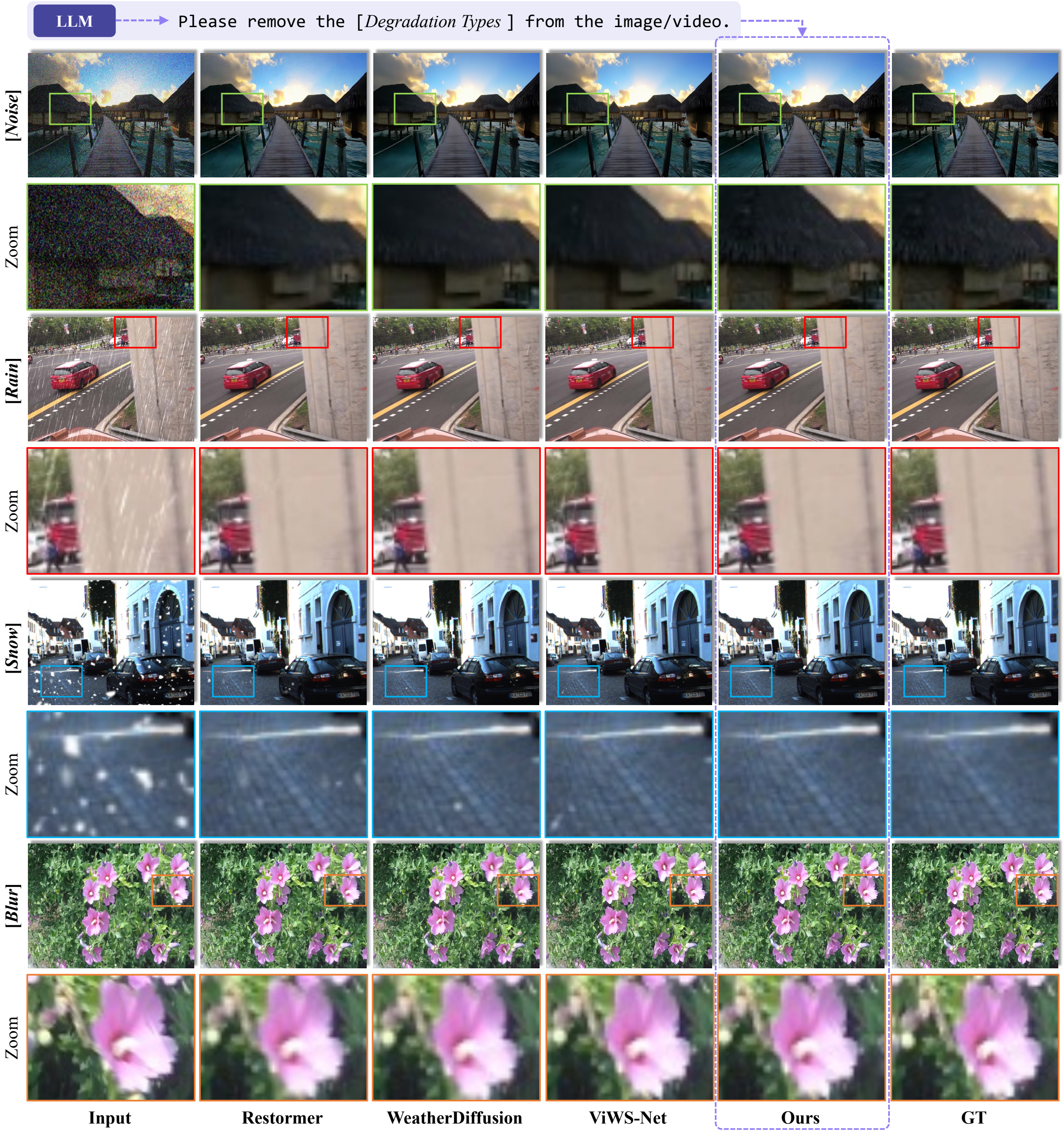


Fig. 4: Qualitative comparisons against state-of-the-art methods under various degradation conditions (noise, rain, snow, and blur). Red boxes indicate regions of interest (ROIs). Best viewed with zoom.

various image degradation scenarios, including Denoise, Derain, Desnow, and Deblur tasks. Specifically, compared to existing methods, our approach better preserves the structural details of the boardwalk in the denoising task, more accurately recovers vehicle contours in the deraining scenario, exhibits stronger restoration capability in nighttime desnowing scenes, and achieves better texture detail preservation of flowers in the deblurring task. In comparison with Ground Truth, it is evident that our method achieves the most visually appealing results in degradation scenarios.

C. Applications in Medical Image & Video Understanding

To further expand the application scenarios of GSDiff and maximize its potential, we extend our video reconstruction framework to medical video and image processing. Video enhancement and reconstruction are vital for clinical diagnosis and treatment planning, particularly in procedures that rely on high-quality imaging or video data, such as endoscopy and surgical procedure [52]. However, the considerable domain gap between medical videos and natural scene footage presents unique challenges when generalizing such models across ap-

TABLE II: Ablation studies on the proposed GSDiff framework. Grid splicing analysis, component analysis and loss function studies validate the effectiveness of our full model design. Higher PSNR and SSIM values indicate better performance. The best values are bolded.

A Impact of Grid Size

Grid Size	Average Performance	
	PSNR↑	SSIM↑
4 (2 × 2)	31.03	0.9289
9 (3 × 3)	**32.33**	**0.9485**
16 (4 × 4)	31.75	0.9422

B Impact of Model Components

Model Strategy	Average Performance	
	PSNR↑	SSIM↑
DBG	25.85	0.6433
DBG + DPM	27.67	0.7121
DBG + DPM + TRN	**32.33**	**0.9485**

C Impact of Loss Terms

Loss Function	Average Performance	
	PSNR↑	SSIM↑
$\mathcal{L}_{diff}$	26.45	0.7220
$\mathcal{L}_{diff} + \mathcal{L}_{rec}$	31.12	0.9105
$\mathcal{L}_{diff} + \mathcal{L}_{rec} + \mathcal{L}_{con}$	**32.33**	**0.9485**

plications. To validate our model's generalizability, we conduct experiments using two challenging medical datasets. First, we select frames from the Kvasir-Capsule dataset [53] and introduce motion blur to simulate the image degradation encountered during endoscopic procedures. Blurring frames are common problems in endoscopy. Factors such as sudden movement, time constraint, organ movement, poor lighting conditions and hand-eye coordination of the operator can all contribute to a noisy or blurry frame, which can hinder following evaluation and diagnosis. Therefore, deblurring is an essential task for accurate and efficient endoscopy.

Additionally, we performed denoising experiments on the low-dose CT dataset [54]. While low dose CT imaging is a safer and more efficient procedure than normal CT, low dose inevitably introduces significant noise artifacts into the CT images, which interfere with diagnosis process and reduce diagnostic accuracy. Therefore, denoising has substantial clinical significance in low-dose CT. Furthermore, we provide qualitative comparisons with representative restoration methods, including Restormer and diffusion-based WeatherDiffusion, demonstrating the superior generalization ability of our approach on medical scenarios.

In Fig. 5, the top rows present performance on low-dose CT images, where our approach effectively suppresses noise artifacts while maintaining anatomical details in both high-contrast lung regions and low-contrast liver areas (red boxes). The bottom rows demonstrate superior deblurring results on endoscopic video frames, where the proposed method preserves intricate tissue structures and subtle textures (green boxes) critical for clinical diagnosis, outperforming existing state-of-the-art approaches.

D. Ablation Study

TABLE III: Ablation Study on Transformer Block Distribution Across Stages in Tail Refine Network. Best results are highlighted in bold.

Transformer Block Layout	Average Performance	
	PSNR↑	SSIM↑
$[2, 2, 2, 2]$	30.94	0.9339
$[4, 4, 4, 4]$ (default)	32.33	0.9485
$[8, 8, 8, 8]$	**32.46**	**0.9431**

*1) **Grid Size**:* In the ablation study of Grid Size (Tab. IIa), we investigate the impact of grid configurations on model performance. The results show that when the grid configuration expands from 4(2×2) to 9(3×3), the performance significantly improves (PSNR increases from 31.03dB to 32.33dB, SSIM from 0.9289 to 0.9485), mainly due to the 3×3 configuration's ability to capture richer temporal dependencies. However, further increasing to 16(4×4) leads to performance degradation (PSNR decreases to 31.75, SSIM to 0.9422), as the increased number of frames results in excessive scene variations that exceed the temporal modeling capacity of the U-Net architecture in Stable Diffusion. This finding demonstrates that the 3×3 configuration achieves an optimal balance between scene variation magnitude and temporal information capture.

*2) **Model Components**:* In the ablation study of model components (Tab. IIb), we progressively validate the contribution of each module. With only LLM and Dual Branch Guidance (DBG), the model achieves basic degradation recognition capability but with limited overall performance (PSNR: 25.85dB, SSIM: 0.6433). After incorporating DPM, the performance improves (PSNR: 27.67dB, SSIM: 0.7121). However, due to the inherent limitations of VAE structures, detail information is inevitably lost during the encoding process. To address this issue, we design the TRN module. To verify the effectiveness of TRN, we demonstrate its advantage in detail preservation through frequency domain analysis (Fig. 6). The spectral visualization reveals that TRN significantly enhances the model's ability to retain high-frequency information, which is intuitively reflected in the reconstruction of vase decorative patterns (red boxed region). In contrast, the model without TRN shows notable loss in high-frequency components, leading to degraded detail reconstruction quality. The frequency domain analysis validates that TRN effectively compensates for the detail loss in VAE, leading to substantial performance gains. The TRN proves effective as a detail compensation module from both frequency and spatial perspectives.

*3) **Loss Function**:* In the ablation study of Loss Terms (Tab. IIc), we investigate the impact of different loss function combinations on model performance. Using only the diffusion loss ($\mathcal{L}_{diff}$), the model achieves basic reconstruction capability (PSNR of 26.45dB, SSIM of 0.7220). After incorporating the reconstruction loss ($\mathcal{L}_{rec}$), the performance improves significantly (PSNR increases to 31.12dB, SSIM to 0.9105) due to direct supervision of reconstruction quality. Finally, introducing the contrastive loss ($\mathcal{L}_{con}$) leads to optimal performance (PSNR reaching 32.33dB, SSIM achieving 0.9485) by enhancing the model's ability to distinguish between different types of degradation. These results demonstrate that each loss term plays an essential role in the optimization process.

*4) **Transformer Block Number**:* To investigate the optimal distribution strategy of Transformer blocks number across stages in the Tail Refine Network, we perform ablation experiments on different configurations and choose $[4, 4, 4, 4]$ as

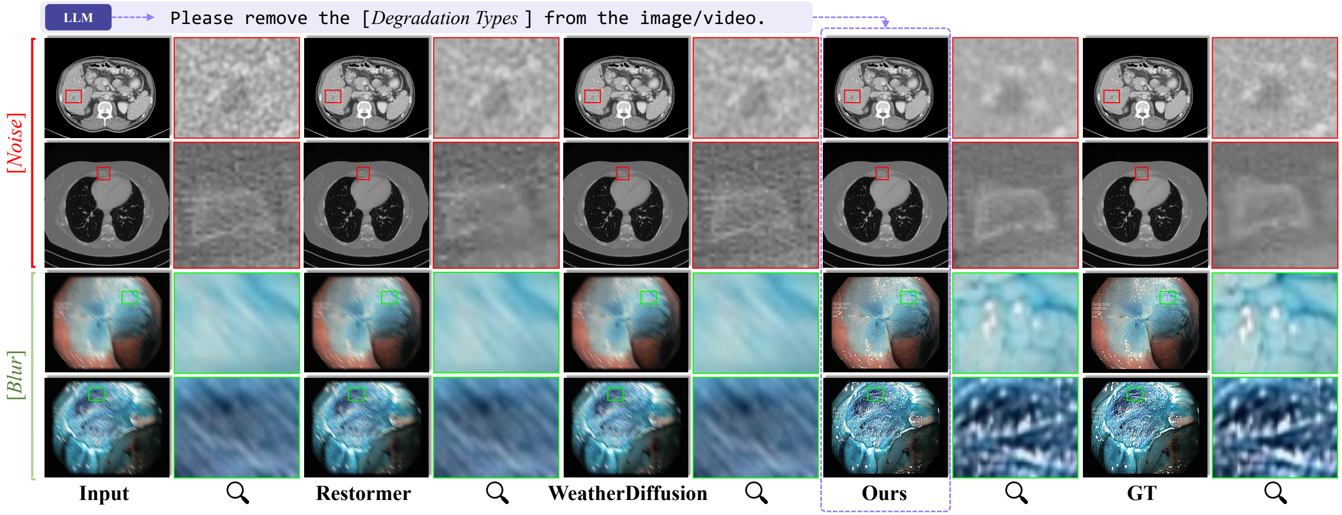

LLM
Please remove the [Degradation Types] from the image/video.
[Noise]
[Blur]
Input
Restormer
WeatherDiffusion
Ours
GT

Fig. 5: Qualitative comparisons on the Low Dose CT dataset for the denoising task (top) and the deblurring task (bottom) on the Kvasir-Capsule dataset. Highlighted ROIs (red and green boxes) demonstrate that our method outperforms other state-of-the-art methods in preserving structural details and textures. Boxes indicate regions of interest (ROIs), best viewed with zoom.

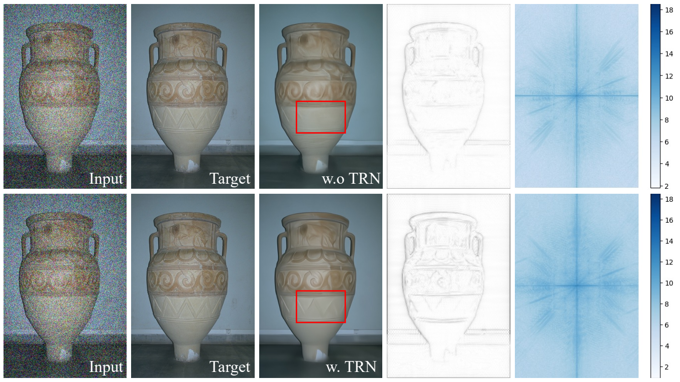


Fig. 6: Ablation study on TRN module. From left to right: input image, target image, outputs without/with TRN, and their corresponding frequency domain visualizations. Red boxes highlight the improved preservation of fine details with TRN.

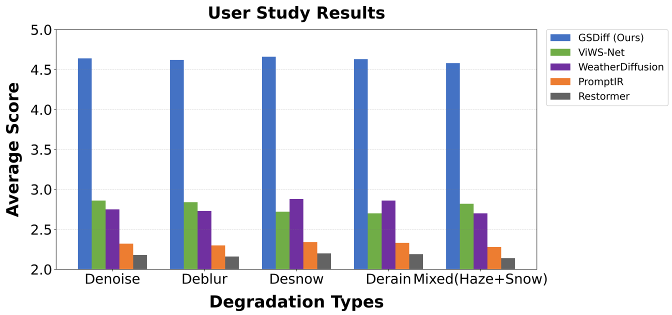


Fig. 7: User study results comparing restoration quality across different degradation types. Our GSDiff achieves superior scores (1-5 scale) in all scenarios.

the default setting for three reasons: First, it ensures balanced feature extraction across all stages, preventing any single stage from becoming a performance bottleneck; Second, empirical evidence indicates that four Transformer blocks are sufficient to effectively model both local and global feature relationships within each stage; Third, this configuration ensures a symmetrical encoder-decoder structure, where the Transformer block number in each encoder stage precisely corresponds to that in the decoder stage, enabling precise one-to-one feature correspondence during the encoding-decoding process for feature reconstruction and restoration. As shown in Tab. III, the shallow $[2, 2, 2, 2]$ configuration performs suboptimally (PSNR: 30.94dB, SSIM: 0.9339) compared with the default $[4, 4, 4, 4]$ setting (PSNR: 32.33dB, SSIM: 0.9485). Although increasing to $[8, 8, 8, 8]$ slightly raises PSNR to 32.46dB, SSIM marginally decreases to 0.9431. Given the substantial computational cost, this improvement provides limited practical value. These experimental results suggest that our default configuration is an optimal choice balancing both model performance and computational efficiency.

E. User Study

To evaluate the effectiveness of our proposed method, we conduct a user study with 50 participants. In our experiments, participants are asked to rate the restoration results (scale 1-5, 5 being the best) across five typical degradation scenarios: denoising, deblurring, desnowing, deraining, and mixed degradation (haze+snow). The comprehensive user study results are presented in Fig. 7, which clearly demonstrates our method's superior performance. Specifically, our GSDiff method achieves significant advantages across all degradation types, maintaining an average score above 4.6 and outperforming the second-best method (ViWS-Net) by approximately 1.8 points. Notably, even in the most challenging mixed degradation scenario, GSDiff maintains a high score of 4.58, which demonstrates the robustness and superior visual restoration quality of our method. The experimental results indicate that our proposed method exhibits excellent generalization capability across various image degradation scenarios.

F. Efficiency comparison

We compare the computational complexity and the efficiency of our parameter choices with other state-of-the-art methods, including TKL [21], WeatherDiffusion [45], and

TABLE IV: Comparison of model parameters and inference speed (best values in bold).

Methods	TKL	WeatherDiffusion	ViWS-Net	GSDiff
Parameters (M)	**28.71**	82.96	57.82	62.36
Inference time (s)	0.51	342.76	**0.46**	0.83

ViWS-Net [31]. Although GSDiff incorporates large pretrained LLM and Stable Diffusion, totaling 5.02B parameters, these pre-trained components remain frozen during both training and inference. For a fair comparison, we only focus on investigating the trainable parameters of each model, as they directly reflect the actual learning capacity and training complexity. As shown in Tab. IV, when considering trainable parameters, TKL demonstrates the most lightweight architecture with 28.71M parameters, followed by ViWS-Net with 57.82M parameters. For GSDiff, while its total parameter count is 5.02B, only the Degradation Extractor, Detail Preservation Module, and Tail Refine Network require additional training, resulting in 62.36M trainable parameters. In terms of inference speed, GSDiff takes 0.83 seconds per iteration, demonstrating competitive efficiency compared to ViWS-Net (0.46s) and TKL (0.51s). In contrast, WeatherDiffusion requires significantly longer inference time at 342.76 seconds. These results highlight that, despite incorporating large pre-trained models, GSDiff maintains unmatched efficiency through its parameter-freezing strategy and adoption of foundation models.

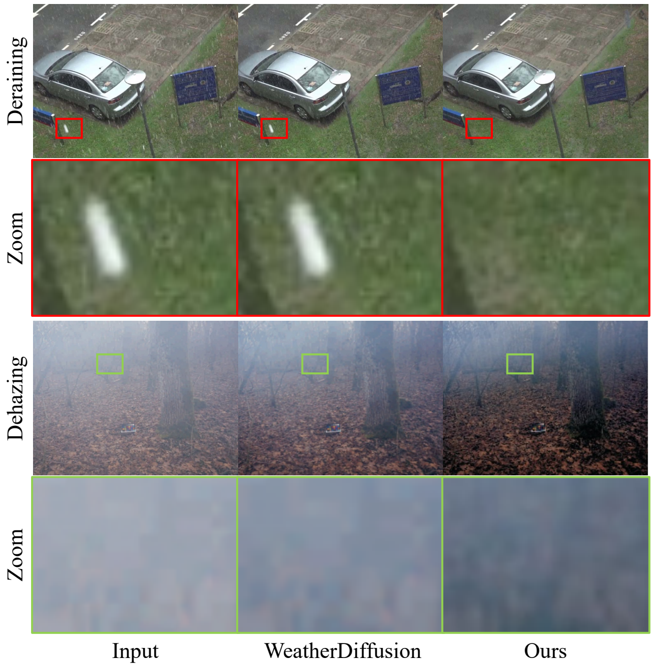


Fig. 8: Visual comparison of real-world weather degradation removal. Top: rain streak removal results with magnified regions. Bottom: defogging results with magnified regions.

G. Real-World Performance Analysis

To validate the effectiveness of our proposed method in complex real-world scenarios, we conduct experiments on two challenging cases: a car-mounted camera video sequence captured in rainy conditions and a forest scene under heavy fog, as shown in Fig. 8. In the rainy scene, the traditional WeatherDiffusion method struggles with noticeable blurring and residual artifacts when processing raindrops. In contrast, our method effectively removes rain interference while preserving image details, particularly in areas like vehicle contours and ground textures.

Tested on the foggy scenes, our algorithm significantly improves over the WeatherDiffusion method. While removing dense fog, our method better preserves scene depth and fine details, avoiding over-smoothing and detail loss. The zoomed-in regions highlight its superiority in reconstructing tree outlines and ground textures. These experimental results demonstrate that our method's enhanced ability to retain details and depth information enables its deployment in real-world applications such as autonomous driving and video surveillance.

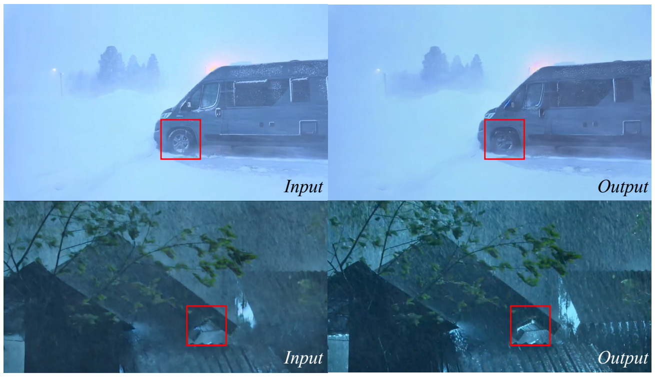


Fig. 9: Visualization of GSDiff limitations under extreme weather conditions. Top: Performance degradation in severe snowstorm. Bottom: Model uncertainty in compound adverse conditions (darkness, snow, and strong winds).

H. Limitation

Despite GSDiff's overall effectiveness, it has limitations when dealing with extreme weather conditions. As illustrated in Fig. 9, our method faces challenges in scenarios with severe degradation. For example, in Fig. 9 (top), we pinpoint a corner case where GSDiff struggles with compound weather conditions involving darkness, snow, and strong winds. In this scenario, GSDiff misclassifies then incorrectly removes street lighting as precipitation artifacts. These challenging cases encourages further refinements in two folds. First, the inherent gap between synthetic training data and real-world scenarios remains a challenge for robust generalization. Second, extreme conditions with multiple severe degradation factors require careful attention and additional studies. These insights provide valuable directions for extending our method, such as advanced multimodal domain adaptation strategies.

V. Conclusion

In this work, we introduce GSDiff, a novel framework for video reconstruction leveraging the synergy between LLMs and diffusion-based generation. Beyond traditional task-specific approaches, our framework demonstrates that integrating advanced AI capabilities from zero-shot recognition to spatiotemporal modeling can lead to more generalized and robust restoration systems. The superior performance across diverse scenarios, from natural scenes to medical imaging, validates our core hypothesis that a unified framework can effectively handle complex real-world degradations without compromising on detail preservation. More importantly, this work opens up new research directions at the intersection of foundation models and low-level vision tasks, suggesting the potential of leveraging semantic understanding to guide low-level feature restoration. Looking forward, we envision this approach evolving beyond mere restoration towards intelligent scene understanding and reconstruction, particularly in mission-critical applications where both visual quality and semantic accuracy are essential. Future research should focus on bridging the gap between controlled and extreme conditions, developing efficient architectures for real-time applications, and exploring domain-specific adaptations while maintaining generalizability. This work represents a step towards next-generation video restoration systems that truly understand and adapt to the complexities of real-world visual degradation.

References

[1] K. C. K. Chan, S. Zhou, X. Xu, and C. C. Loy, "Basicvsr++: Improving video super-resolution with enhanced propagation and alignment," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022.* IEEE, 2022, pp. 5962–5971. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.00588

[2] W. Lai, J. Huang, O. Wang, E. Shechtman, E. Yumer, and M. Yang, "Learning blind video temporal consistency," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11219. Springer, 2018, pp. 179–195. [Online]. Available: https://doi.org/10.1007/978-3-030-01267-0_11

[3] J. Wang, P. Wang, G. Sun, D. Liu, S. A. Dianat, R. Rao, M. Rabbani, and Z. Tao, "Text is MASS: modeling as stochastic embedding for text-video retrieval," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024.* IEEE, 2024, pp. 16 551–16 560. [Online]. Available: https://doi.org/10.1109/CVPR52733.2024.01566

[4] L. Yan, C. Han, Z. Xu, D. Liu, and Q. Wang, "Prompt learns prompt: Exploring knowledge-aware generative prompt collaboration for video captioning," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China.* ijcai.org, 2023, pp. 1622–1630. [Online]. Available: https://doi.org/10.24963/ijcai.2023/180

[5] Z. Qin, X. Lu, X. Nie, D. Liu, Y. Yin, and W. Wang, "Coarse-to-fine video instance segmentation with factorized conditional appearance flows," *IEEE CAA J. Autom. Sinica*, vol. 10, no. 5, pp. 1192–1208, 2023. [Online]. Available: https://doi.org/10.1109/JAS.2023.123456

[6] J. Zhao, Z. Dai, P. Xu, and L. Ren, "Protoviewer: Visual interpretation and diagnostics of deep neural networks with factorized prototypes," in *2020 IEEE Visualization Conference (VIS)*, 2020, pp. 286–290.

[7] Q. Guo, J. Sun, F. Juefei-Xu, L. Ma, X. Xie, W. Feng, Y. Liu, and J. Zhao, "Efficientderain: Learning pixel-wise dilation filtering for high-efficiency single-image deraining," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021.* AAAI Press, 2021, pp. 1487–1495. [Online]. Available: https://doi.org/10.1609/aaai.v35i2.16239

[8] D. Ren, W. Shang, P. Zhu, Q. Hu, D. Meng, and W. Zuo, "Single image deraining using bilateral recurrent network," *IEEE Trans. Image Process.*, vol. 29, pp. 6852–6863, 2020. [Online]. Available: https://doi.org/10.1109/TIP.2020.2994443

[9] T. Ye, S. Chen, Y. Liu, Y. Ye, J. Bai, and E. Chen, "Towards real-time high-definition image snow removal: Efficient pyramid network with asymmetrical encoder-decoder architecture," in *Computer Vision - ACCV 2022 - 16th Asian Conference on Computer Vision, Macao, China, December 4-8, 2022, Proceedings, Part III*, ser. Lecture Notes in Computer Science, L. Wang, J. Gall, T. Chin, I. Sato, and R. Chellappa, Eds., vol. 13843. Springer, 2022, pp. 37–51. [Online]. Available: https://doi.org/10.1007/978-3-031-26313-2_3

[10] P. Li, M. Yun, J. Tian, Y. Tang, G. Wang, and C. Wu, "Stacked dense networks for single-image snow removal," *Neurocomputing*, vol. 367, pp. 152–163, 2019. [Online]. Available: https://doi.org/10.1016/j.neucom.2019.07.023

[11] C. Tian, M. Zheng, W. Zuo, B. Zhang, Y. Zhang, and D. Zhang, "Multi-stage image denoising with the wavelet transform," *Pattern Recognit.*, vol. 134, p. 109050, 2023. [Online]. Available: https://doi.org/10.1016/j.patcog.2022.109050

[12] C. Tian, Y. Xu, W. Zuo, B. Du, C. Lin, and D. Zhang, "Designing and training of a dual CNN for image denoising," *Knowl. Based Syst.*, vol. 226, p. 106949, 2021. [Online]. Available: https://doi.org/10.1016/j.knosys.2021.106949

[13] B. Luo, Z. Cheng, L. Xu, G. Zhang, and H. Li, "Blind image deblurring via superpixel segmentation prior," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1467–1482, 2022. [Online]. Available: https://doi.org/10.1109/TCSVT.2021.3074799

[14] Y. Zhang, Q. Li, M. Qi, D. Liu, J. Kong, and J. Wang, "Multi-scale frequency separation network for image deblurring," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 10, pp. 5525–5537, 2023. [Online]. Available: https://doi.org/10.1109/TCSVT.2023.3259393

[15] Y. Zhang, L. Wei, Q. Zhang, Y. Song, J. Liu, H. Li, X. Tang, Y. Hu, and H. Zhao, "Stable-makeup: When real-world makeup transfer meets diffusion model," *CoRR*, vol. abs/2403.07764, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2403.07764

[16] G. Wang, J. C. Ye, and B. D. Man, "Deep learning for tomographic image reconstruction," *Nat. Mach. Intell.*, vol. 2, no. 12, pp. 737–748, 2020. [Online]. Available: https://doi.org/10.1038/s42256-020-00273-z

[17] L. Zhai, Y. Wang, S. Cui, and Y. Zhou, "A comprehensive review of deep learning-based real-world image restoration," *IEEE Access*, vol. 11, pp. 21 049–21 067, 2023. [Online]. Available: https://doi.org/10.1109/ACCESS.2023.3250616

[18] J. M. J. Valanarasu, R. Yasarla, and V. M. Patel, "Transweather: Transformer-based restoration of images degraded by adverse weather conditions," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022.* IEEE, 2022, pp. 2343–2353. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.00239

[19] R. Li, R. T. Tan, and L. Cheong, "All in one bad weather removal using architectural search," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020.* Computer Vision Foundation / IEEE, 2020, pp. 3172–3182. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Li_All_in_One_Bad_Weather_Removal_Using_Architectural_Search_CVPR_2020_paper.html

[20] O. Özdenizci and R. Legenstein, "Restoring vision in adverse weather conditions with patch-based denoising diffusion models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 10 346–10 357, 2023. [Online]. Available: https://doi.org/10.1109/TPAMI.2023.3238179

[21] W. Chen, Z. Huang, C. Tsai, H. Yang, J. Ding, and S. Kuo, "Learning multiple adverse weather removal via two-stage knowledge learning and multi-contrastive regularization: Toward a unified model," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022.* IEEE, 2022, pp. 17 632–17 641. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.01713

[22] A. Kulkarni, P. W. Patil, S. Murala, and S. Gupta, "Unified multi-weather visibility restoration," *IEEE Trans. Multim.*, vol. 25, pp. 7686–7698, 2023. [Online]. Available: https://doi.org/10.1109/TMM.2022.3225712

[23] Y. Liu, Z. Ke, F. Liu, N. Zhao, and R. W. H. Lau, "Diff-plugin: Revitalizing details for diffusion-based low-level tasks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024.* IEEE, 2024, pp. 4197–4208. [Online]. Available: https://doi.org/10.1109/CVPR52733.2024.00402

[24] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," *CoRR*, vol. abs/2204.06125, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2204.06125

[25] C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang, and N. Duan, "Visual chatgpt: Talking, drawing and editing with visual foundation

models," *CoRR*, vol. abs/2303.04671, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.04671

[26] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 10 674–10 685. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.01042

[27] J. Zhao, X. Liu, C. Guo, Z. C. Qian, and Y. V. Chen, " Phoenixmap: An Abstract Approach to Visualize 2D Spatial Distributions ," *IEEE Transactions on Visualization & Computer Graphics*, vol. 27, no. 03, pp. 2000–2014, Mar. 2021. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/TVCG.2019.2945960

[28] J. Zhao, X. Wang, J. Zhu, C. Chukwudi, A. Finebaum, J. Zhang, S. Yang, S. He, and N. Saeidi, "PhaseFIT: live-organoid phase-fluorescent image transformation via generative AI," *Light Sci Appl*, vol. 12, no. 1, p. 297, Dec 2023.

[29] B. Li, X. Liu, P. Hu, Z. Wu, J. Lv, and X. Peng, "All-in-one image restoration for unknown corruption," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 17 431–17 441. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.01693

[30] V. Potlapalli, S. W. Zamir, S. H. Khan, and F. S. Khan, "Promptir: Prompting for all-in-one image restoration," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023. [Online]. Available: http://papers.nips.cc/paper_files/paper/2023/hash/e187897ed7780a579a0d76fd4a35d107-Abstract-Conference.html

[31] Y. Yang, A. I. Avilés-Rivero, H. Fu, Y. Liu, W. Wang, and L. Zhu, "Video adverse-weather-component suppression network via weather messenger and adversarial backpropagation," in *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 2023, pp. 13 154–13 164. [Online]. Available: https://doi.org/10.1109/ICCV51070.2023.01214

[32] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *CoRR*, vol. abs/2005.14165, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[33] OpenAI, "GPT-4 technical report," *CoRR*, vol. abs/2303.08774, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.08774

[34] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, "Palm: Scaling language modeling with pathways," *J. Mach. Learn. Res.*, vol. 24, pp. 240:1–240:113, 2023. [Online]. Available: https://jmlr.org/papers/v24/22-1144.html

[35] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," *CoRR*, vol. abs/2302.13971, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2302.13971

[36] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 2022, pp. 12 888–12 900. [Online]. Available: https://proceedings.mlr.press/v162/li22n.html

[37] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. C. H. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023. [Online]. Available: http://papers.nips.cc/paper_files/paper/2023/hash/9a6a435e75419a836fe47ab6793623e6-Abstract-Conference.html

[38] T. Wang, Y. Liu, J. C. Liang, J. Zhao, Y. Cui, Y. Mao, S. Nie, J. Liu, F. Feng, Z. Xu, C. Han, L. Huang, Q. Wang, and D. Liu, "M^{2}PT: Multimodal prompt tuning for zero-shot instruction learning," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 3723–3740. [Online]. Available: https://aclanthology.org/2024.emnlp-main.218/

[39] B. Zheng, J. Gu, S. Li, and C. Dong, "Lm4lv: A frozen large language model for low-level vision tasks," *arXiv preprint arXiv:2405.15734*, 2024.

[40] S. Sirnam, J. Yang, T. Neiman, M. N. Rizve, S. Tran, B. Yao, T. Chilimbi, and M. Shah, "X-former: Unifying contrastive and reconstruction learning for mllms," in *European Conference on Computer Vision*. Springer, 2024, pp. 146–162.

[41] T. Wang, Y. Liu, J. C. Liang, J. Zhao, Y. Cui, Y. Mao, S. Nie, J. Liu, F. Feng, Z. Xu, C. Han, L. Huang, Q. Wang, and D. Liu, "M^{2}PT: Multimodal prompt tuning for zero-shot instruction learning," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 3723–3740. [Online]. Available: https://aclanthology.org/2024.emnlp-main.218/

[42] Z. Geng, L. Liang, T. Ding, and I. Zharkov, "RSTT: real-time spatial temporal transformer for space-time video super-resolution," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 17 420–17 430. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.01692

[43] R. Li, R. T. Tan, and L. Cheong, "All in one bad weather removal using architectural search," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020, pp. 3172–3182. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Li_All_in_One_Bad_Weather_Removal_Using_Architectural_Search_CVPR_2020_paper.html

[44] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 5718–5729. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.00564

[45] O. Özdenizci and R. Legenstein, "Restoring vision in adverse weather conditions with patch-based denoising diffusion models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[46] Y. Zhu, T. Wang, X. Fu, X. Yang, X. Guo, J. Dai, Y. Qiao, and X. Hu, "Learning weather-general and weather-specific features for image restoration under multiple adverse weather conditions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 747–21 758.

[47] S. Wang, L. Zhu, H. Fu, J. Qin, C. Schönlieb, W. Feng, and S. Wang, "Rethinking video rain streak removal: A new synthesis model and a deraining network with video rain prior," in *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XIX*, ser. Lecture Notes in Computer Science, S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., vol. 13679. Springer, 2022, pp. 565–582. [Online]. Available: https://doi.org/10.1007/978-3-031-19800-7_33

[48] W. Chen, H. Fang, C. Hsieh, C. Tsai, I. Chen, J. Ding, and S. Kuo, "ALL snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 4176–4185. [Online]. Available: https://doi.org/10.1109/ICCV48922.2021.00416

[49] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. V. Gool, M. H. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 724–732. [Online]. Available: https://doi.org/10.1109/CVPR.2016.85

[50] M. Tassano, J. Delon, and T. Veit, "Fastdvdnet: Towards real-time deep video denoising without flow estimation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020.* Computer Vision Foundation / IEEE, 2020, pp. 1351–1360. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Tassano_FastDVDnet_Towards_Real-Time_Deep_Video_Denoising_Without_Flow_Estimation_CVPR_2020_paper.html

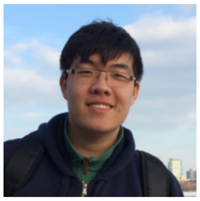
[51] S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017.* IEEE Computer Society, 2017, pp. 257–265. [Online]. Available: https://doi.org/10.1109/CVPR.2017.35

[52] S. Huang, Y. Ge, D. Liu, M. Hong, J. Zhao, and A. C. Loui, "Rethinking copy-paste for consistency learning in medical image segmentation," *IEEE Transactions on Image Processing*, vol. 34, pp. 1060–1074, 2025.

[53] K. Pogorelov, H. K. Stensland, D.-T. Dang-Nguyen, M. Gamnes, V. Thambawita, M. Riegler, and P. Halvorsen, "Kvasir-capsule, a video capsule endoscopy dataset," *Scientific Data*, vol. 8, no. 1, p. 142, 2021.

[54] C. H. McCollough, A. C. Bartley, R. E. Carter, B. Chen, T. A. Drees, P. Edwards, D. R. Holmes, A. E. Huang, F. Khan, S. Leng, K. L. McMillan, G. J. Michalak, K. M. Nunez, L. Yu, and J. G. Fletcher, "Low-dose CT for the detection and classification of metastatic liver lesions: Results of the 2016 Low Dose CT Grand Challenge," *Medical Physics*, vol. 44, no. 10, pp. e339–e352, 2017.

Jinliang Liu is currently pursuing the Ph.D. degree with the University of Technology Sydney, Australia. He received the M.S. degree from the Australian National University, Australia, in 2019. His research interests include image and video generation, multimodal learning and AutoML.



Jianwei Zang is a Ph.D. student at University of Southern California. He worked after Aydogan Ozcan and received a M.S. degree from University of California, Los Angeles (UCLA). Prior to this, he received a B.S. degree from Carnegie Mellon University in computer engineering. His research is focused on deep learning applications in brain computer interface system.

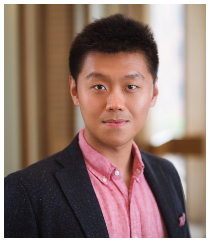
Sen Yang is a life science research scientist at Stanford University. He received the B.E. and M.S. degrees in 2017 and 2020 from Sichuan University, Chengdu, China. His research interest includes machine learning and deep learning.

Jinxi Xiang is a postdoctoral scholar at Stanford University. He is a multidisciplinary researcher specializing in signal processing and machine learning for healthcare applications. Xiang integrated machine learning with medical imaging during my doctoral studies at Tsing-Hua University. As a scientist at Tencent AI Lab, he developed AI tools for clinical pathology and gaming, focusing on image/video coding and multimodal learning.

Xiyue Wang Xiyue Wang received the B.E., M.S., and Ph.D. degrees from Sichuan University, Chengdu, China, in 2017, 2020, and 2023, respectively. Currently, she is a Post-Doctoral Researcher with Stanford University, focusing on medical image analysis and personalized cancer treatment. Her research interests include analyzing pathological images, with a specific emphasis on developing generalized self-supervised feature representations and exploring the potential of weakly supervised learning methods.

Jieqiong Zhao is currently a Tenure-Track Assistant Professor in the School of Computer and Cyber Sciences at Augusta University. Prior to this, she was a postdoc research associate in the VADER lab at the School of Computing and Augmented Intelligence (SCAI), Arizona State University, working under the guidance of Dr. Ross Maciejewski. She received my Ph.D. in Electrical and Computer Engineering from Purdue University, where I was mentored by Dr. David S. Ebert, and a master's degree in Computer Science from Tufts University, under the mentorship of Dr. Remco Chang.

Zongxin Yang received his Bachelor's degree in Engineering (BE) from the University of Science and Technology of China in 2018 and earned his Ph.D. in Computer Science from the University of Technology Sydney in 2021. He is currently a postdoctoral researcher at Harvard University. His research interests include multi-modal learning, vision generation, and the intersection of biomedical science and AI.



Junhan Zhao is a Fellow at Harvard Medical School, specializing in AI-driven digital health. Zhao earned his Ph.D. in computer graphics as a Bilsland Fellowship awardee and served as a lecturer at Purdue University. He completed his engineering training and graduated with distinction from Shanghai Jiao Tong University and Cornell University. He also holds a M.Sc in Biostatistics from Harvard University. Zhao has served as first or senior author on high-impact publications featured in Nature, Med, Light Science & Applications, and IEEE TVCG, TIP, TMI, TIM and TCSVT. He led AI development and deployment in early-stage ventures.